DOI: 10.1093/nar/gkh426

MotifViz: an analysis and visualization tool for motif discovery

Yutao Fu[1], Martin C. Frith[1], Peter M. Haverty[1] and Zhiping Weng[1,2,*]

[1]Bioinformatics Program and [2]Biomedical Engineering Department, Boston University, Boston, MA 02215, USA

Received February 15, 2004; Revised March 30, 2004; Accepted April 12, 2004

ABSTRACT

Detecting overrepresented known transcription factor binding motifs in a set of promoter sequences of co-regulated genes has become an important approach to deciphering transcriptional regulatory mechanisms. In this paper, we present an interactive web server, MotifViz, for three motif discovery programs, Clover, Rover and Motifish, covering most available flavors of algorithms for achieving this goal. For comparison, we have also implemented the simple motif-matching program Possum. MotifViz provides uniform and intuitive input and output formats for all four programs. It can be accessed at http://biowulf.bu.edu/MotifViz.

INTRODUCTION

With the rapid accumulation of genomic sequence data and the development of high-throughput experimental technologies such as DNA microarrays and their combination with chromatin immunoprecipitation, recent years have witnessed a substantial increase in computational efforts toward the understanding of transcriptional regulation. The identification of short sequence motifs, such as transcription factor binding sites, is at the center of such efforts. Regulatory motifs are often overrepresented in the promoters of co-regulated genes, which can be detected by microarray studies of genetically altered or chemically stimulated cells. There are two main approaches toward finding such motifs: (i) *ab initio* motif discovery with multiple local sequence alignment and (ii) the detection of statistical overrepresentation of previously known motifs. While much attention has been paid to *ab initio* algorithms [e.g. (1–5)], the detection of overrepresentation is simpler and potentially more powerful, because the search is confined to a library of known motifs (6). However, the latter approach is not able to identify previously unknown motifs.

Algorithms for testing whether a motif is overrepresented in a target set of DNA sequences have appeared only recently (6–12). Most of these methods ultimately reduce to the statistics of contingency tables. A position-specific scoring matrix (PSSM) for a motif is first scanned across the target sequences and a set of control sequences, and matches with a similarity score greater than some threshold are recorded. The data can be cast as a 2×2 contingency table, with the four entries of the table denoting (i) the number of *target* sequences *with* a match, (ii) the number of *control* sequences *with* a match, (iii) the number of *targets without* a match and (iv) the number of *controls without* a match. A χ^2 test or Fisher's exact test (using the hypergeometric distribution) can then be performed to test the null hypothesis that the sequences with motif matches are evenly distributed among the target and control sets. Three flavors of such *sequence-based* methods have been published (7–9). Since target and control sequences are typically of different lengths and can contain different numbers of motifs, several *motif-based* methods have been proposed to count the total number of motifs rather than sequences, and construct a similar contingency table to that described above (8,10–12). We have recently developed a method named Clover which cannot be classified as sequence- or motif-based. Rather, it combines multiple matches per sequence in an intuitive manner, motivated by a simple thermodynamic model (6). We further establish the statistical significance of the results via comparison to sequences obtained by randomizing nucleotides or dinucleotides in the target sequences, by randomizing the columns of each PSSM and by selecting random segments from a large set of background DNA sequences (6).

With the accumulation of known motifs in databases such as TRANSFAC (13) and JASPAR (14), motif-overrepresentation algorithms are becoming sufficiently accurate for guiding laboratory experimentation. Active collaborations centered on the application of such algorithms have sprung up between our lab and three experimental labs studying estrogen response elements, dopamine response genes and platelet-specific genes. We quickly realized the need for a user-friendly web interface for motif-overrepresentation algorithms. Here we describe an interactive web server for a motif-based algorithm Rover (12), a sequence-based algorithm Motifish (M. C. Frith,

*To whom all correspondence should be addressed. Tel: +1 617 353 3509; Fax: +1 617 353 6766; Email: zhiping@bu.edu

Y. Fu and Z. Weng, unpublished results) and Clover (6). For comparison, we also include a simple program named Possum which scans one sequence against one motif, since this represents the base-line prediction.

MotifViz makes these four algorithms available to a much wider audience and facilitates the comparison of results from different approaches. This web server provides a focused toolset for motif detection, in contrast to other applications that are only available as command-line driven and computer-platform-specific programs, or are buried in large software packages. With experimental scientist users in mind, we have carefully designed a consistent web interface that gives a similar look and feel to all four algorithms. Both an overview of motif distribution and a detailed sequence output are presented, with cross-referencing and mouse-over functions to facilitate the selection of sequence regions for experimental testing. The common input and output formats we provide for this spectrum of motif detection algorithms allow the user to identify the relevant benefits of different methods and focus later research on higher likelihood results agreed upon by multiple methods.

DESCRIPTION OF MotifViz

Methods

Possum scans one motif against a sequence and calculates the log-likelihood ratio score for the motif at every location of the sequence:

$$\text{Score} = \log\left(\prod_{k=1}^{W} \frac{q(k, L_k)}{p(L_k)}\right), \qquad \mathbf{1}$$

where W is the width of the motif, L denotes the location being considered, L_k is the nucleotide at position k within this location and $p(X)$ is the background probability of observing nucleotide X, estimated from the frequency of X in that sequence. $q(k, X)$ is the probability of nucleotide X at position k in the motif. This score is the standard motif matrix score, motivated by a simple thermodynamic model (15,16).

Motifish takes as input two sets of sequences, the target set and the control set (ideally, they should have equal length and similar base composition). Taking one motif at a time, Motifish counts the number of sequences with one or more motif matches (with the score defined in Equation 1 above a user-defined threshold) in the target and control sets. The P-value for overrepresentation of counts in the target set is calculated using Fisher's exact test.

Clover and Rover have been described in detail previously (6,12). It is important to point out the differences between these methods. Clover, Rover and Motifish scan through a library of motifs and detect those that are overrepresented in target versus control sequences. Possum simply finds individual motifs, and does not evaluate overrepresentation. Therefore, Possum makes a large number of predictions, and should only be applied to several selected motifs (e.g. the motifs predicted by the other three methods). Motifish performs a rapid and simple test of motif overrepresentation, but it requires all input sequences to be of the same length, and ignores the presence of multiple motif hits in a sequence. Clover integrates multiple sites in the same sequence and multiple-site-containing sequences in a more mathematically sophisticated way than Rover. However, Clover is the slowest method. The user is encouraged to run all three methods; the motifs predicted by multiple methods may have a higher chance of being biologically functional.

Input

All four programs require two items of input: a set of target DNA sequences and a selection of motifs. Rover and Motifish require a set of background (or control) DNA sequences. Clover provides the option to take background sequences or, alternatively, it can compute P-values from sequence or motif shuffling. Each program also has a small number of adjustable parameters. We have designed the MotifViz web interface to integrate all four programs seamlessly. The user is asked to pick a program first, and only the input and parameter items pertinent to the chosen program will be displayed. Shared items are placed at the top of the web page and in the same arrangement.

Sequences may be entered in FASTA, raw or GenBank formats. Promoter sequences corresponding to transcripts can be easily obtained from the PromoSer web service (17). Alternatively, the user can supply GenBank identifiers and MotifViz will automatically fetch the sequences. Any non-alphabetic characters in the input sequences will be ignored, and any alphabetic characters except A, C, G and T (uppercase or lowercase) will be converted to Ns and excluded from matching motifs. More sophisticated treatments of ambiguous bases are possible, but would provide minuscule improvements in sensitivity at the cost of specificity and program complexity. MotifViz will read and display any CDS (protein-coding region) annotations in GenBank format. A maximum of 50 target sequences is allowed, and the total length of the sequences is limited to 100 kb.

Motifs can be entered as matrices, where rows correspond to sequential positions in the motif and columns indicate abundances of A, C, G and T at each position. Similar to the FASTA format for sequences, each matrix is preceded by a title line which consists of a '>' sign followed by the motif name. In addition to user-defined motifs, we provide the JASPAR collection of eukaryotic motifs (14). The user can pick a subset of the JASPAR motifs, or select motif families. Another source of motifs is TRANSFAC, which is a commercial database (free registration for academic use) (13). MotifViz recognizes TRANSFAC's motif format and the user can directly copy and paste matrices from the TRANSFAC website. These two motif formats can be mixed.

The use of background sequence sets to contrast with sequences of interest is required for Rover and Motifish, and is highly recommended for P-value evaluation with Clover. Each user-provided background file should contain sequences in FASTA format. Clover requires that the total length of the background sequences be much greater than the target sequence set. Both ROVER and Motifish can also benefit from large background sets. However, larger sequence sets will increase computation time. We provide six background sets for Clover, and multiple background sets can be used for this program. The choice of background is extremely important. Ideally, background sequences should posses exactly the same characteristics as target sequences but lack the motifs enriched in target sequences. Specifically, the two sequence sets should come from the same taxonomic group, contain similar repetitive elements and have similar compositional biases (e.g. similar GC content, dinucleotide or higher-order nucleotide composition). When target sequences are promoters of co-regulated genes obtained from a microarray experiment, promoters of genes with invariant expression according to the same experiment could constitute a good control set, since they are most likely not to contain motifs of interest.

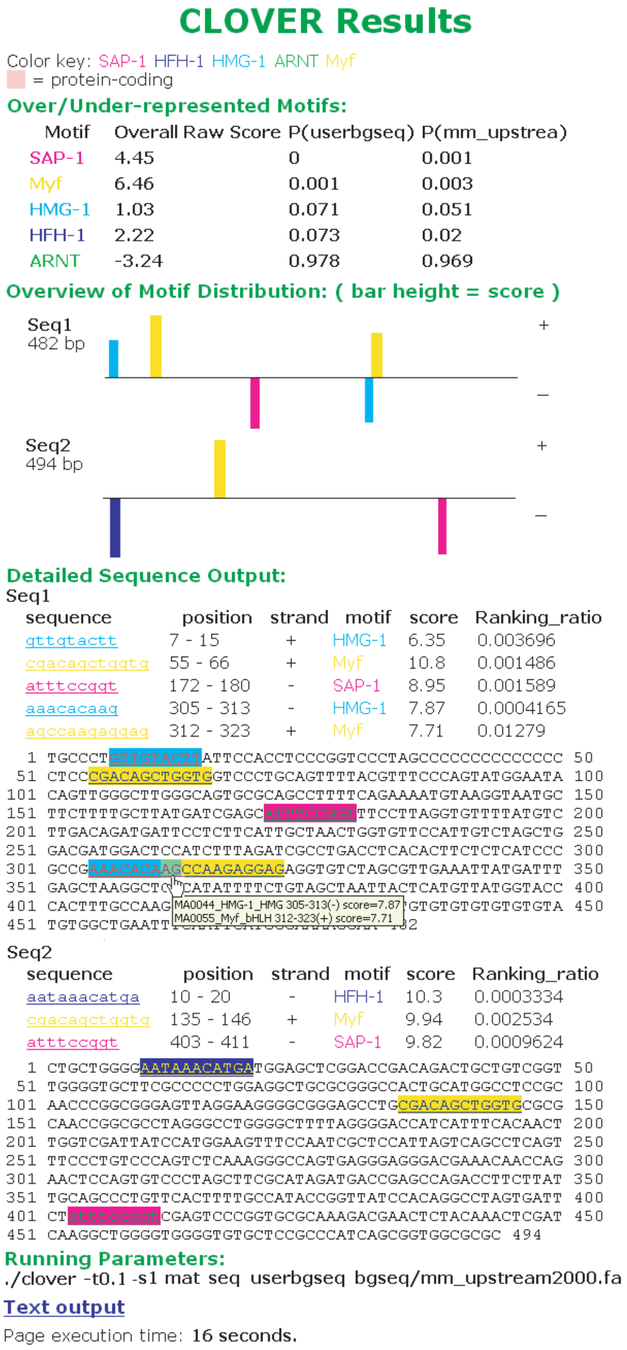
There are a number of parameters that can affect each program's behavior. Possum reports all motif hits that score above a threshold according to Equation 1. Possum estimates the background probability of each nucleotide [$p(X)$ in Equation 1] by counting nucleotide abundances in a window centered on that nucleotide. The size of this window is adjustable. All four programs add pseudocounts to all entries in the motif matrices, and the number of pseudocounts is an adjustable parameter. The addition of pseudocounts is a widely used technique for estimating underlying frequencies from a limited number of counts, with a theoretical underpinning in Bayesian statistics. All programs can filter sequence regions in lowercase letters, which is becoming a standard way of indicating repetitive elements.

Clover, Rover and Motifish compute P-values to indicate the degree of overrepresentation of each PSSM in target versus control sequence sets. Clover uses randomization tests to obtain P-values, and the user may specify the number of randomizations to be performed. Larger numbers of shuffles can provide more accurate P-values at the expense of computation time. Clover has three cutoffs for reporting purposes: a P-value threshold, an overall raw score threshold and a score threshold for individual motif hits. When P-values are computed, all motifs with P-values below the threshold are reported. Otherwise, motifs with overall raw scores above the raw score threshold are reported. In the detailed output (see below), only motif hits that score above the individual motif score threshold are reported. Rover and Motifish provide a similar P-value threshold for reporting motifs. In addition, these two programs require a cutoff to qualify a site as a motif hit. For this purpose, the user can specify the frequency of motif hits in the background set for Rover, and the percentage of background sequences containing the motif for Motifish.

Output

Output appears in a new browser window to allow the user to adjust input parameters and compare different programs. Separate motif types are colored differently using a scheme that selects colors from evenly separated locations in the rainbow spectrum. Motifs from the same family are assigned similar colors. Figure 1 is an example of Clover output obtained by searching two mouse sequences for all motifs in JASPAR, using mouse promoter sequences as background. Output files for Clover, Rover and Motifish contain three parts: motif report, graphical overview and detailed sequence information. Possum output contains only the latter two parts. The first part lists discovered motifs along with their overall raw scores and P-values that measure the degrees of overrepresentation.

The second part of the output provides a graphical overview of the distribution of the motifs within each input sequence. A



CLOVER Results

Color key: SAP-1 HFH-1 HMG-1 ARNT Myf
▨ = protein-coding

Over/Under-represented Motifs:

Motif	Overall Raw Score	P(userbgseq)	P(mm_upstrea)
SAP-1	4.45	0	0.001
Myf	6.46	0.001	0.003
HMG-1	1.03	0.071	0.051
HFH-1	2.22	0.073	0.02
ARNT	-3.24	0.978	0.969

Overview of Motif Distribution: (bar height = score)

Detailed Sequence Output:

Seq1

sequence	position	strand	motif	score	Ranking_ratio
gttgtactt	7 - 15	+	HMG-1	6.35	0.003696
cgacagctggtg	55 - 66	+	Myf	10.8	0.001486
atttccggt	172 - 180	-	SAP-1	8.95	0.001589
aaacacaag	305 - 313	-	HMG-1	7.87	0.0004165
agccaagaggag	312 - 323	+	Myf	7.71	0.01279

```
  1 TGCCCTGTTGTACTTATTCCACCTCCCGGTCCCTAGCCCCCCCCCCCCCC 50
 51 CTCCCGACAGCTGGTGGTCCCTGCAGTTTTACGTTTCCCAGTATGGAATA 100
101 CAGTTGGGCTTGGGCAGTGCGCAGCCTTTTCAGAAAATGTAAGGTAATGC 150
151 TTCTTTTGCTTATGATCGAGCATTTCCGGTTTCCTTAGGTGTTTTATGTC 200
201 TTGACAGATGATTCCTCTTCATTGCTAACTGGTGTTCCATTGTCTAGCTG 250
251 GACGATGGACTCCATCTTTAGATCGCCTGACCTCACACTTCTCTCATCCC 300
301 GCCGAAACACAAGCCAAGAGGAGAGGTGTCTAGCGTTGAAATTATGATTT 350
351 GAGCTAAGGCTC CATATTTTCTGTAGCTAATTACTCATGTTATGGTACC 400
401 CACTTTGCCAAG MA0044_HMG-1_HMG 305-313(-) score=7.87 GTGTGTGTGTGTGTA 450
451 TGTGGCTGAATT MA0055_Myf_bHLH 312-323(+) score=7.71 82
```

Seq2

sequence	position	strand	motif	score	Ranking_ratio
aataaacatga	10 - 20	-	HFH-1	10.3	0.0003334
cgacagctggtg	135 - 146	+	Myf	9.94	0.002534
atttccggt	403 - 411	-	SAP-1	9.82	0.0009624

```
  1 CTGCTGGGGAATAAACATGATGGAGCTCGGACCGACAGACTGCTGTCGGT 50
 51 TGGGGTGCTTCGCCCCCTGGAGGCTGCGCGGGCCACTGCATGGCCTCCGC 100
101 AACCCGGCGGGAGTTAGGAAGGGGCGGGAGCCTGCGACAGCTGGTGCGCG 150
151 CAACCGGCGCCTAGGGCCTGGGGCTTTTAGGGGACCATCATTTCACAACT 200
201 TGGTCGATTATCCATGGAAGTTTCCAATCGCTCCATTAGTCAGCCTCAGT 250
251 TTCCCTGTCCCAGTCTCAAAGGGCCAGTGAGGGAGGGACGAAACAACCAG 300
301 AACTCCAGTGTCCCTAGCTTCGCATAGATGACCGAGCCAGACCTTCTTAT 350
351 TGCAGCCCTGTTCACTTTTGCCATACCGGTTATCCACAGGCCTAGTGATT 400
401 CTatttccggtCGAGTCCCGGTGCGCAAAGACGAACTCTACAAACTCGAT 450
451 CAAGGCTGGGGTGGGGTGTGCTCCGCCCATCAGCGGTGGCGCGC 494
```

Running Parameters:

```
./clover -t0.1 -s1 mat seq userbgseq bgseq/mm_upstream2000.fa
```

Text output

Page execution time: **16 seconds.**

Figure 1. Example output of Clover. The user-supplied background sequence and a mouse background sequence supplied by MotifViz are used for P-value calculation. The overall raw scores and P-values are reported for all over/underrepresented motifs in the top section of the output. Only two target sequences are used for illustrative purpose. Plus and minus signs indicate the motif hits in the positive and negative strands of the sequences. In the second section, an overview of all hits of overrepresented motifs in each sequence is illustrated with color bars, with bar height corresponding to raw score and bar width corresponding to motif width. In the third section, a detailed report is given for all motif hits in the target sequences. For each motif hit, in addition to the raw score computed according to Equation 1, we also report a P-value-like measure called ranking ratio, which indicates the percentage of sites that score better than the raw score in a large segment of the human genome. Mousing over a motif hit invokes a popup window indicating the name, location and raw score of the hit (this is illustrated with the overlapping HMG-1 and Myf sites in the first sequence). At the end of the output, a hyperlink is provided for saving the text output of this run.

horizontal line represents the sequence and vertical color bars indicate locations of the corresponding motifs. Bars below the line represent occurrence of motifs on the opposite strand. Bar height represents the score of each motif occurrence. The Perl GD graphics library interface (by Lincoln D. Stein, Cold Spring Harbor Laboratory) is used for image creation.

The third part of the output is a detailed text printout grouped by sequences. The exact locations (sequence, strand, and start and end positions), motif names and scores are listed in a table. In order to provide a *P*-value-like measure for individual motif hits, we have scanned each JASPAR PSSM against 2000-bp upstream sequences of all genes in the human genome, and reported the percentage of sites that score above a specified threshold. This percentage is called ranking ratio and is reported alongside the raw scores of all individual motif hits, which are used as the thresholds during the calculation of ranking ratios. Hits that score below 0 are excluded during the scan to save disk space; therefore, the ranking ratios for such hits are 100%. As a result, our ranking ratio is more conservative than the actual *P*-value; however, the discrepancy is negligible for high-scoring motif hits, which are the hits of biological interest. Overlapping motif sites are represented by blending the colors of the corresponding motifs. A popup window reporting detailed information about the motif (same as in the table) is invoked by passing the mouse cursor over a highlighted motif. The highlighted motif sites in a sequence and those in the table are cross-referenced with each other, i.e. a user can move the mouse cursor back and forth between the two locations by clicking on either site. Such an interactive display of the output can greatly facilitate the navigation among many predicted motifs and the extraction of sub-sequences for experimental testing.

Visualization of command-line outputs

When a user saves a MotifViz output page as an HTML file, some old web browsers may not save the images automatically. The text output, to which we provide a web link at the end of the output page, can be saved as a text file and uploaded to MotifViz to reproduce the graphical output. When large query sequences or multiple large background sequences are used for any program, or various sequence/motif shuffling approaches are combined for Clover, a job may take a long time to finish. To eliminate the wait for such time-consuming jobs, as well as to reduce the burden on our web server, the computational workload should be shifted to the user's local machines. The user can download all four programs, run them locally from the command line, and save the output as a plain-text file. The file then can be uploaded to MotifViz for visualization.

ACKNOWLEDGEMENTS

We are grateful to Katya Ravid and Jun Lu for motivation and helpful discussions on output format. This work was partly supported by NSF Major Research Instrument grant DBI-0116574. Z.W. and M.C.F. were supported in part by NSF grant DBI-0078194. Z.W. was partially funded by NIH grant 1P20GM066401-01. Y.F. and Z.W. were partially funded by NIH grant 1R01HG03110-01. P.M.H. was supported by NSF grant IGERT-9870710. M.C.F. was a Howard Hughes Medical Institute Predoctoral Fellow.

REFERENCES

1. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
2. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol*., **2**, 28–36.
3. Hughes,J.D., Estep,P.W., Tavazoie,S. and Church,G.M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol*., **296**, 1205–1214.
4. Liu,X., Brutlag,D.L. and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput*., 127–138.
5. Frith,M.C., Hansen,U., Spouge,J.L. and Weng,Z. (2004) Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res*., **32**, 189–200.
6. Frith,M.C., Fu,Y., Yu,L., Chen,J.F., Hansen,U. and Weng,Z. (2004) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res*., **32**, 1372–1381.
7. Liu,R., McEachin,R.C. and States,D.J. (2003) Computationally identifying novel NF-kappa B-regulated immune genes in the human genome. *Genome Res*., **13**, 654–661.
8. Sharan,R., Ovcharenko,I., Ben-Hur,A. and Karp,R.M. (2003) CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics*, **19**(Suppl 1), I283–I291.
9. Elkon,R., Linhart,C., Sharan,R., Shamir,R. and Shiloh,Y. (2003) Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res*., **13**, 773–780.
10. Aerts,S., Thijs,G., Coessens,B., Staes,M., Moreau,Y. and Moor,B.D. (2003) Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res*., **31**, 1753–1764.
11. Zheng,J., Wu,J. and Sun,Z. (2003) An approach to identify over-represented cis-elements in related sequences. *Nucleic Acids Res*., **31**, 1995–2005.
12. Haverty,P.M., Hansen,U. and Weng,Z. (2004) Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. *Nucleic Acids Res*., **32**, 179–188.
13. Wingender,E., Kel,A.E., Kel,O.V., Karas,H., Heinemeyer,T., Dietze,P., Knuppel,R., Romaschenko,A.G. and Kolchanov,N.A. (1997) TRANSFAC, TRRD and COMPEL: towards a federated database system on transcriptional regulation. *Nucleic Acids Res*., **25**, 265–268.
14. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*., **32**, D91–D94.
15. Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol*., **193**, 723–750.
16. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
17. Halees,A.S., Leyfer,D. and Weng,Z. (2003) PromoSer: A large-scale mammalian promoter and transcription start site identification service. *Nucleic Acids Res*., **31**, 3554–3559.